Building an annotated ark

Genome sequencing technologies are improving, and researchers have their eyes on a lot more animals.

Michael Eisenstein

Two by two? Ambitious projects to sequence thousands of species are filling up the 'ark' of animal genome sequences. Credit: E. Dewalt / Springer Nature

Something was wrong with the monkeys. Animal care staff at the California National Primate Research Center noted that several young rhesus macaques were having trouble adapting to changes in their housing. "They seemed to do fine when they were in their home cage," explains Jeffrey Rogers, a geneticist at Baylor College of Medicine. "But when they moved to a new cage, the monkeys were literally feeling their way around." Suspecting a congenital vision problem, the Center sought help from Rogers, who previously helped coordinate the sequencing of the macaque genome[1].

The researchers ultimately homed in on a disruptive mutation in a gene that encodes a light-responsive protein in the cone cells of the retina[2]. Their findings not only explain the animals' eyesight issues but also offer a promising new model for achromatopsia, a disorder that causes poor color perception and reduced visual acuity but has proven tough to study in conventional mouse models. "There are fundamental differences between the visual system in a nocturnal mouse and a diurnal human," Rogers says. But with a detailed rhesus macaque genome as a reference, he and his Baylor colleague Rui Chen were able to home in on a human-relevant mutation in the macaques.

More novel models are likely waiting to be found. Rapid evolution in hardware and software for DNA analysis and falling costs per experiment are making it easier for scientists to prospect the genomes of classic model organisms as well as novel species that intrigue them. Some groups are using this approach to explore biomedical questions in species with characteristics that parallel human traits, as seen with studies of cancer and behavioral disorders in domesticated dogs or vocal communication in songbirds. Others are studying species with unusual features that might nevertheless prove beneficial to human health, such as long-lived but cancer- and virus-resistant bats or the highly regenerative axolotl.

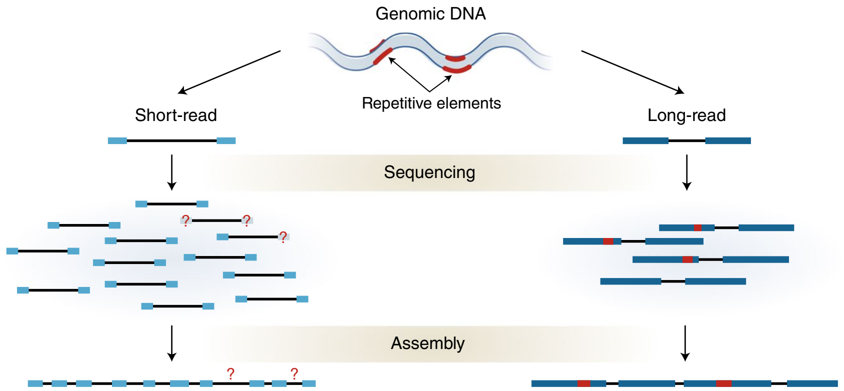
Genomicists are also banding together to achieve even loftier goals. These include efforts to corral full genomes for every species of bird and bat or, more ambitiously, for all 66,000 vertebrates—the objective of the multinational Vertebrate Genomes Project (VGP), coordinated by Erich Jarvis at Rockefeller University. Paul Flicek, who studies vertebrate genomics at the European Molecular Biology Laboratory's European Bioinformatics Institute (EBI), is among those who believe that the resulting datasets will provide an invaluable resource for uncovering the molecular foundations of many of life's mysteries. "Evolution leaves behind patterns, and computationally you can go and look and find those patterns," he says.

> "I would love to get to a point—and I don't know how possible this is—where we can make predictions based on genome sequence and functional information as to what a given species might be a good model for," says Paul Flicek

A better read

Soon after the publication of the 'first draft' of the human genome in 2000, biologists began contemplating similar efforts for some of the most widely used model species—in particular, the laboratory mouse, *Mus musculus*. But having a clear roadmap from the Human Genome Project did not make this follow-up simple or cheap. It took the Mouse Genome Sequencing Consortium years of work at a cost of millions of dollars to reconstruct the 2.5 billion base-pairs, or gigabases, of DNA that make up the mouse genome, through a technique called 'shotgun' analysis[3]. "You just cut the genome into many small pieces and sequence enough of them that you can put the puzzle together," says Kerstin Lindblad-Toh, a scientific director at the Broad Institute who was one of the leaders of this sequencing effort. The order of nucleotides in these fragments was determined through an arduous but effective biochemical procedure called Sanger sequencing, one of the earliest methods of DNA sequence analysis. Sanger sequencing, originally developed in the late 1970s, also allowed researchers to tackle other important animal genomes, including the chimpanzee and dog.

However, it wasn't until the late 2000s that animal genomics truly picked up momentum. In 2009, the biotechnology company Illumina began marketing a sequencing platform that could rapidly generate billions upon billions of short ‘reads’ from a given DNA sample in a single experiment. Although tiny, spanning just a couple hundred nucleotides each, the sheer abundance and remarkable accuracy of these reads dramatically accelerated the rate at which genome assemblies could be completed. For a well-studied genome like ours or that of the mouse, contemporary Illumina instruments can spit out a full sequence in just over a day. Illumina’s technology now dominates the sequencing world, but it‘s not a total solution for the *de novo*—that is, from scratch—assembly of novel genomes. Such assemblies break down in the long stretches of highly-repetitive DNA that punctuate many genomes, where reads spanning just a few hundred bases can yield ambiguous data that is hard to interpret. “By 2010 or 2011, we could generate reasonably complete pictures of most genes, but not very contiguous assemblies,” explains Benedict Paten, a computational genomicist at the University of California at Santa Cruz.



Getting a read on it: Short-read sequencing systems work by generating millions of short sequence ‘reads’ spanning a few hundred bases that are then computationally aligned based on their overlap to reconstruct the original sequence. Long-read strategies reduce the number of pieces needed by assembling the sequence from reads that can span thousands of bases. Long reads can contain more errors than short reads but can produce more accurate genomes because of their more extensive overlap and ability to encompass long stretches of repetitive DNA. Repetitive elements in short reads can be ambigous to interpet and lead to gaps in the assembly. Credit: E. Dewalt / Springer Nature

> “Think of two jigsaw puzzles of the same picture, where one is in 30,000 pieces and the other is in four,” says Emma Teeling.

A subsequent wave of technological innovation brought this task within reach, with the rise of ‘long-read’ sequencing platforms that generate continuous strings spanning tens of thousands of bases. This produces greater overlap and coverage of the genome, making assembly far easier. “Think of two jigsaw puzzles of the same picture, where one is in 30,000 pieces and the other is in four,” says Emma Teeling, who studies the molecular evolution of bats at University College Dublin. The advantage of long reads has become particularly apparent with challenging species such as the axolotl, a Mexican salamander with remarkable powers of limb regeneration, whose genome had previously proven insurmountable. “You have something like 18 gigabases of repetitive elements, and the elements that have expanded the most often run longer than 10 kilobases—so you need long reads to span them,” says Michael Hiller of the Max Planck Institute of Molecular Cell Biology and Genetics, who collaborated on the axolotl genome[4].

Many genomics efforts now employ long-read sequencing technology developed by Pacific Biosciences (PacBio), which is relatively costly but can generate extremely high-quality genome assemblies. More recently, a company called Oxford Nanopore Technologies has offered a more affordable platform for long-read sequencing, and several researchers are now exploring this alternative. “We’re currently assembling the cynomolgus macaque genome using only Oxford Nanopore reads,” says Rogers.

From Jarvis’s point of view, quality and accuracy are utmost priorities. This is the product of frustrations encountered in his efforts to study evolution of vocal learning in birds. “When we started digging down into genomes that were being produced, we discovered that there were a lot of errors that made it difficult to answer these questions confidently,” he says. To obtain ‘platinum-grade’ quality data, he and his VGP collaborators are combining strategies, using short reads to ‘fact check’ the long-read data as well as other techniques that can bridge even larger contiguous assemblies—in effect, further reducing the number of pieces in the puzzle. For example, a technique called ‘Hi-C’ can reveal whether two independent fragments are situated on the same chromosome, allowing researchers to build continuous sequences spanning tens of millions of bases.

This isn’t cheap—even with negotiated discounts, Jarvis’s team spends $15,000 to assemble a modest 1 gigabase genome—but there are hidden costs to more economical approaches. “People would rather spend $2–3,000 on a short-read assembly, but they don’t realize how much work and time students and postdocs will have to spend to correct all the mistakes,” he says.

Assembly line

A flawlessly executed sequencing experiment can provide all the right pieces to a genome puzzle, but it still takes careful processing to ensure they are fitted together correctly. Unfortunately, this is not as straightforward as one might hope. “Software can make mistakes where there are no mistakes in the raw data,” says Jarvis. Guojie Zhang, co-organizer of the ‘Bird 10,000 Genomes’ project, notes that the cost and time required for computational analysis routinely exceeds that of the sequencing itself.

Highly repetitive sequences remain one of the greatest enemies of accurate assembly. Long reads mitigate the problem, but do not eliminate it—particularly for large and complicated genomes like those of sharks or amphibians. Jarvis notes that in a pilot run for the VGP’s genomics workflow, procedures that performed very well in birds and mammals fell flat with the skate,

Go long: Assembling the axolotl’s genome meant dealing with a highly repetitive sequence 10x longer than that of a human. Credit: Kouichi Tsunoda / EyeEm / Getty

Bats up: The Bat1K Project's goal is to sequence every bat species, including the greater mouse eared-bat *Myotis myotis*. Credit: Oliver Farcy

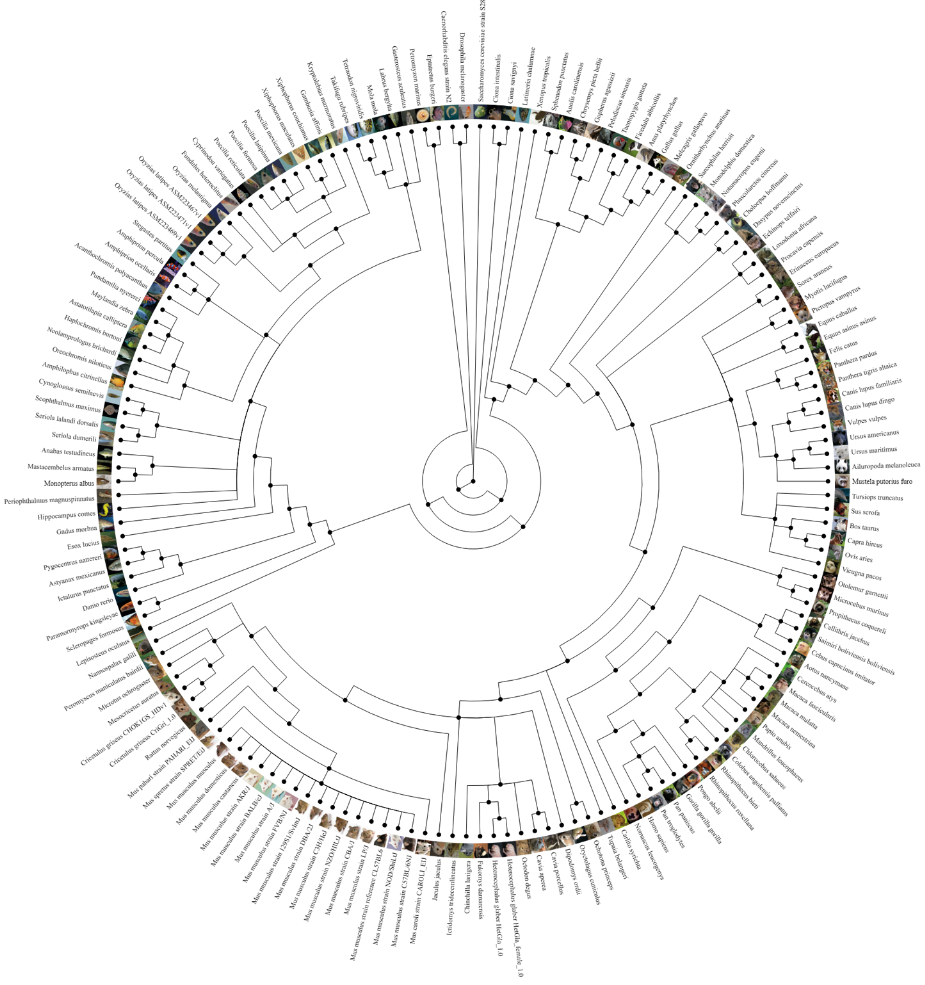
whose genome is nearly twice the length of the human genome and roughly 50% repetitive sequences. Hiller had a similar experience in wrangling the axolotl genome, which weighs in at a whopping 32 billion base-pairs—ten times larger than the human genome—with existing software. "Pretty much everything that typically works stops working with genomes of that size," he says. This necessitated the creation of specialized tools that could cope with such data.

Even for simpler genomes, like those of mammals, assembly is not necessarily a cake-walk. Mammals are generally diploid, meaning they carry two copies of each chromosome—one from each parent. Each chromosome copy carries numerous sequence variants relative to its counterpart, and these should ideally be 'phased' during assembly. This means that variants residing on the same chromosome copy are also linked in the reconstructed sequence. If not, two gene copies that are actually 'twins' might be misinterpreted as 'cousins'—additional versions of a gene produced by duplication events that arose through historical errors in DNA replication. Such errors can seriously confound analysis, because real gene duplication events are both common and critical in evolutionary history. "Much of vertebrate evolution involves a pattern of genome duplication, followed by further functionalization," says Paten.

The next task is to fill in where the various genes reside on the emerging map. This is typically achieved with a combination of manual curation and software tools to annotate new genomes. Sequences that actually produce proteins are relatively easy to identify at this point in time. "We know the 'grammar' of genes and the genetic code, and so we can predict which mutations change genes or their encoded proteins," says Hiller. Many genomics efforts also employ a parallel strategy known as 'transcriptomics', using sequencing systems to catalogue the RNA output of a cell or tissue—data that can then be matched back to the genome to reveal the genes from which each RNA originated.

Unfortunately, protein-coding sequences are a minority relative to the far vaster terrain of functional but non-coding sequences, including regulatory elements with a powerful influence on gene expression. These are not well-catalogued and their defining features are often poorly understood, and so charting them remains a far harder—but still essential—task. "That's where the magic lies—it's the 'dark matter' of the genome," says Teeling.

Fortunately, each new annotated genome can simplify the analysis of those that come after. "It's easy for us to borrow evidence from nearby species, because there's relatively little evolutionary change even over five or ten million years in most cases," says Flicek, who leads the Ensembl project, one of the leading efforts in genome collection and annotation. More distant relatives may not be as helpful for annotation, but alignments between such species can nevertheless be deeply informative for addressing questions of evolutionary history—particularly among large numbers of diverse species. Finding alignments is a computationally demanding task, but Paten notes that distributed 'cloud' computing strategies are now making it feasible to comb through a myriad of genomes in parallel in search of sources of commonality and divergence—a major goal of efforts like the VGP. "This allows us to basically spin up thousands of compute nodes and compute at the scale of one or two or even five million CPU-hours to generate one of these alignments," he says. Using such comparisons, Hiller's team was able to track down specific changes in developmental gene regulation that caused snakes to lose their limbs[5].



A growing Ensembl: The Ensembl project is sequencing genomes and building tools, like this visual representation of the relationships between species, for understanding the growing number of genomic datasets. Credit: Ensembl

Box 1 | VGP in progress

The VGP plans to complete 260 genomes in 2019. In their first data release last fall, they published genomes for the following animals:

Mammals
- Greater horseshoe bat - *Rhinolophus ferrumequinum*
- Spear-nosed bat - *Phyllostomus discolor*
- Canadian lynx - *Lynx canadensis*
- Platypus - *Ornithorhynchus anatinus*

Birds
- Anna's hummingbird - *Calypte anna*
- Zebra finch (male & female) - *Taeniopygia guttata*
- Kakapo - *Strigops habroptilus*

Reptiles
- Goode's desert tortoise - *Gopherus evgoodei*

Amphibians
- Two-lined caecilian - *Rhinatrema bivittatum*

Fishes
- Flier cichlid - *Archocentrus centrarchus*
- Eastern happy - *Astatotilapia calliptera*
- Climbing perch - *Anabas testudineus*
- Tire track eel - *Mastacembelus armatus*
- Blunt-snouted clingfish - *Gouania willdenowi*

New model army

These are still early days—Flicek estimates that genome sequences are only available for less than 1% of known vertebrate species, and only a fraction of these meet today's high quality standards. But progress is coming rapidly—the VGP aims to complete its first batch of 260 species within the coming year (Box 1), and several groups are racing to flesh out their favorite limbs of the tree of life. Teeling is coordinating a project called Bat1K, which aims to collect high-quality sequences for all 1,300 bat species within five years. Rogers projects a similar time-frame for mapping out humanity's closest relatives. "We're probably going to have whole-genome sequences for at least one individual from each genus of primates, and probably multiple species from most of them," he says. And although vertebrates are getting the lion's share of attention, a number of parallel efforts are also underway in the invertebrate world—including the Global Ant Genomics Alliance, coordinated by Zhang and colleagues, which aims to gather genomic data for at least 200 different ant species. "All these species have different social structures and caste systems, live in different environments, and have unique behaviors," says Zhang.

These sequencing efforts are already yielding exciting findings that would be hard or impossible to obtain with rodent models. For example, Lindblad-Toh is exploring the genetic foundations of diseases common to particular dog breeds in order to home in on factors involved in human health. In one such effort, her team identified five genes that likely contribute to obsessive-compulsive disorder (OCD) in dogs, as well as four OCD-associated genes that act in similar pathways in humans[6]. "One was the serotonin receptor 2A, which is interesting because one method of treatment for dogs and humans is serotonin reuptake inhibitors," she says. "It would be interesting to see if mutations in that gene correlate with response to that drug."

Part of Teeling's inspiration for embarking on Bat 1K was curiosity about the remarkable longevity and healthspan of many bat species. She has already begun to glean valuable insights from some of the longer-lived specimens sequenced to date. "These bats are living to the equivalent of 258 human years based on their body size with little to no sign of aging," she says. Genome data so far have revealed a number of potential contributors, including a pair of genes that apparently work overtime to preserve chromosomal stability well into old age[7], and several tiny RNAs that appear to keep a tight lid on inflammation[8]—a major culprit in age-related disease and tissue degeneration. Bats may have evolved stronger expression of these RNAs to contain the damaging effects of the metabolic hyperactivity associated with flight, but Teeling notes that humans also have genes encoding equivalent RNAs. "It may be we could use those little switches in ourselves," she says.

These genomes are also a powerful adjunct for research into animal species that were already the subject of intense scientific interest—for example, initial assessment of the axolotl genome has offered starting points for uncovering the biological foundation of tissue regeneration. Perhaps even more exciting is the possibility that scientists could actively discover previously unrecognized candidate models that match their research interests. "I would love to get to a point—and I don't know how possible this is—where we can make predictions based on genome sequence and functional information as to what a given species might be a good model for," says Flicek. Not every animal is ideal for experimental work, like Teeling's long-lived and slowly reproducing bats, but genome-guided tissue culture experiments from these species could nevertheless be deeply informative. Genetic engineering techniques like CRISPR editing could enable the transplantation and functional study of distinctive traits and genes into more amenable models, including mice.

And above all, the assembled genomes promise to serve as a true atlas of life's diversity—the 21st century's answer to the naturalists' sketchbooks and museum specimens of yore. "Rather than having a thing in a jar, you're going to have a digital sequence," says Teeling. "You'll have a true representation of that species at this time on this planet—and this will be used as a reference for all posterity." □

Michael Eisenstein
Freelance science writer, Philadelphia, PA, USA.
e-mail: michael@eisensteinium.com

Published online: 18 March 2019
https://doi.org/10.1038/s41684-019-0270-5

References

1. Rhesus Macaque Genome Sequencing and Analysis Consortium. *Science* **316**, 222–243 (2007).
2. Moshiri, A. et al. *J Clin Invest.* **129**, 863–874 (2019).
3. Mouse Genome Sequencing Consortium. *Nature* **420**, 520–562 (2002).
4. Nowoshilow, S. et al. *Nature* **554**, 50–55 (2018).
5. Roscito, J. G. et al. *Nat. Comm.* **9**, 4737 (2018).
6. Tang, R. et al. *Genome Biol.* **15**, R25 (2014).
7. Foley, N. M. et al. *Sci. Adv.* **4**, eaao0926 (2018).
8. Huang, Z. et al. *BMC Genomics* **17**, 906 (2016).